

# Cross-Lingual Syntactic Transfer with Limited Resources

Mohammad Sadegh Rasooli and Michael Collins\*

Department of Computer Science, Columbia University  
New York, NY 10027, USA  
{rasooli, mcollins}@cs.columbia.edu

## Abstract

We describe a simple but effective method for cross-lingual syntactic transfer of dependency parsers, in the scenario where a large amount of translation data is not available. The method makes use of three steps: 1) a method for deriving cross-lingual word clusters, that can then be used in a multilingual parser; 2) a method for transferring lexical information from a target language to source language treebanks; 3) a method for integrating these steps with the density-driven annotation projection method of Rasooli and Collins (2015). Experiments show improvements over the state-of-the-art in several languages used in previous work (Rasooli and Collins, 2015; Zhang and Barzilay, 2015; Ammar et al., 2016), in a setting where the only source of translation data is the Bible, a considerably smaller corpus than the Europarl corpus used in previous work. Results using the Europarl corpus as a source of translation data show additional improvements over the results of Rasooli and Collins (2015). We conclude with results on 38 datasets (26 languages) from the Universal Dependencies corpora: 13 datasets (10 languages) have unlabeled attachment accuracies of 80% or higher; the average unlabeled accuracy on the 38 datasets is 74.8%.

## 1 Introduction

Creating manually-annotated syntactic treebanks is an expensive and time consuming task. Recently there has been a great deal of interest in cross-lingual

syntactic transfer, where a parsing model is trained for some language of interest, using only treebanks in other languages. There is a clear motivation for this in building parsing models for languages for which treebank data is unavailable. Methods for syntactic transfer include annotation projection methods (Hwa et al., 2005; Ganchev et al., 2009; McDonald et al., 2011; Ma and Xia, 2014; Rasooli and Collins, 2015; Lacroix et al., 2016; Agić et al., 2016), learning of delexicalized models on universal treebanks (Zeman and Resnik, 2008; McDonald et al., 2011; Täckström et al., 2013; Rosa and Zabokrtsky, 2015), treebank translation (Tiedemann et al., 2014; Tiedemann, 2015; Tiedemann and Agić, 2016) and methods that leverage cross-lingual representations of word clusters, embeddings or dictionaries (Täckström et al., 2012; Durrett et al., 2012; Duong et al., 2015; Zhang and Barzilay, 2015; Xiao and Guo, 2015; Guo et al., 2015; Guo et al., 2016; Ammar et al., 2016).

This paper considers the problem of cross-lingual syntactic transfer with limited resources of monolingual and translation data. Specifically, we use the Bible corpus of Christodouloupoulos and Steedman (2014) as a source of translation data, and Wikipedia as a source of monolingual data. We deliberately limit ourselves to the use of Bible translation data because it is available for a very broad set of languages: the data from Christodouloupoulos and Steedman (2014) includes data from 100 languages. The Bible data contains a much smaller set of sentences (around 24,000) than other translation corpora, for example Europarl (Koehn, 2005), which has around 2 million sentences per language

---

\*On leave at Google Inc. New York.

pair. This makes it a considerably more challenging corpus to work with. Similarly, our choice of Wikipedia as the source of monolingual data is motivated by the availability of Wikipedia data in a very broad set of languages.

We introduce a set of simple but effective methods for syntactic transfer, as follows:

- We describe a method for deriving cross-lingual clusters, where words from different languages with a similar syntactic or semantic role are grouped in the same cluster. These clusters can then be used as features in a shift-reduce dependency parser.
- We describe a method for transfer of lexical information from the target language into source language treebanks, using word-to-word translation dictionaries derived from parallel corpora. Lexical features from the target language can then be integrated in parsing.
- We describe a method that integrates the above two approaches with the density-driven approach to annotation projection described in (Rasooli and Collins, 2015).

Experiments show that our model outperforms previous work on a set of European languages from the Google universal treebank (McDonald et al., 2013): we achieve 80.9% average unlabeled attachment score (UAS) on these languages; in comparison the work of Zhang and Barzilay (2015), Guo et al. (2016) and Ammar et al. (2016) have UAS of 75.4%, 76.3% and 77.8% respectively. All of these previous works make use of the much larger Europarl (Koehn, 2005) corpus to derive lexical representations. When using Europarl data instead of the Bible, our approach gives 83.86% accuracy, a 1.68% absolute improvements over (Rasooli and Collins, 2015). Finally, we conduct experiments on 38 datasets (26 languages) in the universal dependencies v1.3 (Nivre et al., 2016) corpus. Our method has an average unlabeled dependency accuracy of 74.8% for these languages, compared to an average accuracy of 68.1% for the method of Rasooli and Collins (2015). 13 datasets (10 languages) have accuracies higher than 80.0%.<sup>1</sup>

<sup>1</sup> The parser code is available at <https://github.com/rasoolims/YaraParser/tree/transfer>.

## 2 Background

This section gives a description of the underlying parsing models used in our experiments, the data sets used, and a baseline approach based on delexicalized parsing models.

### 2.1 The Parsing Model

We assume that the parsing model is a discriminative linear model, where given a sentence  $x$ , and a set of candidate parses  $\mathcal{Y}(x)$ , the output from the model is

$$y^*(x) = \arg \max_{y \in \mathcal{Y}(x)} \theta \cdot \phi(x, y)$$

where  $\theta \in \mathbb{R}^d$  is a parameter vector, and  $\phi(x, y)$  is a feature vector for the pair  $(x, y)$ . In our experiments we use the shift-reduce dependency parser of Rasooli and Tetreault (2015), which is an extension of the approach in (Zhang and Nivre, 2011). The parser is trained using the averaged structured perceptron (Collins, 2002).

We assume that the feature vector  $\phi(x, y)$  is the concatenation of three feature vectors:

- $\phi^{(p)}(x, y)$  is an unlexicalized set of features. Each such feature may depend on the part-of-speech (POS) tag of words in the sentence, but does not depend on the identity of individual words in the sentence.
- $\phi^{(c)}(x, y)$  is a set of cluster features. These features require access to some dictionary that maps each word in the sentence to an underlying cluster identity. Clusters may for example be learned using the Brown clustering algorithm (Brown et al., 1992). The features may make use of cluster identities in combination with POS tags.
- $\phi^{(l)}(x, y)$  is a set of lexicalized features. Each such feature may depend directly on word identities in the sentence. These features may also depend on part-of-speech tags or cluster information, in conjunction with lexical information.

Appendix A has a full description of the features used in our experiments.

## 2.2 Data Assumptions

Throughout this paper we will assume that we have  $m$  source languages  $\mathcal{L}_1 \dots \mathcal{L}_m$ , and a single target language  $\mathcal{L}_{m+1}$ . We assume the following data sources:

**Source language treebanks.** We have a treebank  $\mathcal{T}_i$  for each language  $i \in \{1 \dots m\}$ .

**Part-of-speech (POS) data.** We have hand-annotated POS data for all languages  $\mathcal{L}_1 \dots \mathcal{L}_{m+1}$ . We assume that the data uses a universal POS set that is common across all languages.

**Monolingual data.** We have monolingual raw data for each of the  $(m+1)$  languages. We use  $\mathcal{D}_i$  to refer to the monolingual data for the  $i$ 'th language.

**Translation data.** We have translation data for all language pairs. We use  $\mathcal{B}_{i,j}$  to refer to translation data for the language pair  $(i, j)$  where  $i, j \in \{1 \dots (m+1)\}$  and  $i \neq j$ .

In our main experiments we use the Google universal treebank (McDonald et al., 2013) as our source language treebanks<sup>2</sup> (this treebank provides universal dependency relations and POS tags), Wikipedia data as our monolingual data, and the Bible data from Christodouloupoulos and Steedman (2014) as the source of our translation data. In additional experiments we use the Europarl corpus as a source of translation data, in order to measure the impact of using the smaller Bible corpus.

## 2.3 A Baseline Approach: Delexicalized Parsers with Self-Training

Given the data assumption of a universal POS set, the feature vectors  $\phi^{(p)}(x, y)$  can be shared across languages. A simple approach is then to simply train a delexicalized parser using treebanks  $\mathcal{T}_1 \dots \mathcal{T}_m$ , using the representation  $\phi(x, y) = \phi^{(p)}(x, y)$  (see (McDonald et al., 2013; Täckström et al., 2013)).

Our baseline approach makes use of a delexicalized parser, with two refinements:

**WALS properties.** We use the six properties from the world atlas of language structures (WALS) (Dryer and Haspelmath, 2013) to select a subset of

Feature	Description
82A	Order of subject and verb
83A	Order of object and verb
85A	Order of adposition and noun phrase
86A	Order of genitive and noun
87A	Order of adjective and noun
88A	Order of demonstrative and noun

Table 1: The six properties from the world atlas of language structures (WALS) (Dryer and Haspelmath, 2013) used to select the source languages for each target language in our experiments.

closely related languages for each target language. These properties are shown in Table 1. The model for a target language is trained on treebank data from languages where at least 4 out of 6 WALS properties are common between the source and target language.<sup>3</sup> This gives a slightly stronger baseline: our experiments showed an improvement in average labeled dependency accuracy for the languages from 62.52% to 63.18%. Table 2 shows the set of source languages used for each target language; these source languages are used for all experiments in the paper.

**Self-training.** We use self-training (McClosky et al., 2006) to further improve parsing performance. Specifically, we first train a delexicalized model on treebanks  $\mathcal{T}_1 \dots \mathcal{T}_m$ ; then use the resulting model to parse a dataset  $\mathcal{T}_{m+1}$  that includes target-language sentences which have POS tags but do not have dependency structures. We finally use the automatically parsed data  $\mathcal{T}'_{m+1}$  as the treebank data and retrain the model; this last model is trained using all features (unlexicalized, clusters, and lexicalized). Self-training in this way gives an improvement in labeled accuracy from 63.18% to 63.91%.

## 2.4 Translation Dictionaries

Our only use of the translation data  $\mathcal{B}_{i,j}$  for  $i, j \in \{1 \dots (m+1)\}$  is to construct a translation dictionary  $t(w, i, j)$ . Here  $i$  and  $j$  are two languages,  $w$  is a word in language  $\mathcal{L}_i$ , and the output  $w' = t(w, i, j)$  is a word in language  $\mathcal{L}_j$  corresponding to the most frequent translation of  $w$  into this language.

<sup>2</sup>We also train our best performing model on the newly released universal treebank v1.3 (Nivre et al., 2016). See §4.3 for more details.

<sup>3</sup>There was no effort to optimize this choice; future work may consider more sophisticated sharing schemes.

Target	Sources
en	de, fr, pt, sv
de	en, fr, pt
es	fr, it, pt
fr	en, de, es, it, pt, sv
it	es, fr, pt
pt	en, de, es, fr, it, sv
sv	en, fr, pt

Table 2: The selected source languages for each target language in the Google universal treebank v2 (McDonald et al., 2013). A language is regarded as source if it has at least 4 out of 6 WALS properties in common with the target language.

We define the function  $t(w, i, j)$  as follows. We first run the GIZA++ alignment process (Och and Ney, 2000) on the data  $\mathcal{B}_{i,j}$ . We then keep intersected alignments between sentences in the two languages. Finally, for each word  $w$  in  $\mathcal{L}_i$ , we define  $w' = t(w, i, j)$  to be the target language word most frequently aligned to  $w$  in the aligned data. If a word  $w$  is never seen aligned to a target language word  $w'$ , we define  $t(w, i, j) = \text{NULL}$ .

Future work may consider the use of probabilistic lexicons, or hand-crafted lexicons, as an alternative to the method described above.

### 3 Our Approach

This section describes our approach on improving over the baseline delexicalized approach. The following subsections describe our approach: §3.1 describes a method for deriving cross-lingual clusters, allowing us to add cluster features  $\phi^{(c)}(x, y)$  to the model. §3.2 describes a method for adding lexical features  $\phi^{(l)}(x, y)$  to the model. §3.3 describes a method for integrating the approach with the density-driven approach of Rasooli and Collins (2015). Finally, §4 describes experiments. We show that each of the above steps leads to improvements in accuracy.

#### 3.1 Learning Cross-Lingual Clusters

We now describe a method for learning cross-lingual clusters. This follows previous work on cross-lingual clustering algorithms (Täckström et al., 2012). A *clustering* is a function  $C(w)$  that maps

---

**Inputs:** 1) Monolingual texts  $\mathcal{D}_i$  for  $i = 1 \dots (m + 1)$ ; 2) a function  $t(w, i, j)$  that translates a word  $w \in \mathcal{L}_i$  to  $w' \in \mathcal{L}_j$ ; and 3) a parameter  $\alpha$  such that  $0 < \alpha < 1$ .

**Algorithm:**

```

 $\mathcal{D} = \{\}$ 
for  $i = 1$  to  $m + 1$  do
  for each sentence  $s \in \mathcal{D}_i$  do
    for  $p = 1$  to  $|s|$  do
      Sample  $\bar{a} \sim [0, 1)$ 
      if  $\bar{a} \geq \alpha$  then
        continue
      Sample  $j \sim \text{unif}\{1, \dots, m + 1\} \setminus \{i\}$ 
       $w' = t(s_p, i, j)$ 
      if  $w' \neq \text{NULL}$  then
        Set  $s_p = w'$ 
   $\mathcal{D} = \mathcal{D} \cup \{s\}$ 

```

Use the algorithm of Stratos et al. (2015) on  $\mathcal{D}$  to learn a clustering  $\mathcal{C}$ .

**Output:** The clustering  $\mathcal{C}$ .

---

Figure 1: An algorithm for learning a cross-lingual clustering. In our experiments we used the parameter value  $\alpha = 0.3$ .

each word  $w$  in a vocabulary to a cluster  $C(w) \in \{1 \dots K\}$ , where  $K$  is the number of clusters. A *hierarchical clustering* is a function  $C(w, l)$  that maps a word  $w$  together with an integer  $l$  to a cluster at level  $l$  in the hierarchy. As one example, the Brown clustering algorithm (Brown et al., 1992) gives a hierarchical clustering. The level  $l$  allows cluster features at different levels of granularity.

A *cross-lingual* hierarchical clustering is a function  $C(w, l)$  where the clusters are shared across the  $(m + 1)$  languages of interest: that is, the word  $w$  can be from any of the  $(m + 1)$  languages. Ideally, a cross-lingual clustering should put words across different languages which have a similar syntactic and/or semantic role in the same cluster. There is a clear motivation for cross-lingual clustering in the parsing context. We can use the cluster-based features  $\phi^{(c)}(x, y)$  on the source language treebanks  $\mathcal{T}_1 \dots \mathcal{T}_m$ , and these features will now generalize beyond these treebanks to the target language  $\mathcal{L}_{m+1}$ .

We learn a cross-lingual clustering by leveraging the monolingual data sets  $\mathcal{D}_1 \dots \mathcal{D}_{m+1}$ , together

with the translation dictionaries  $t(w, i, j)$  learned from the translation data. Figure 1 shows the algorithm that learns a cross-lingual clustering. The algorithm first prepares a multilingual corpus, as follows: for each sentence  $s$  in the monolingual data  $\mathcal{D}_i$ , for each word in  $s$ , with probability  $\alpha$  we replace the word with its translation into some randomly chosen language. Once this data is created, we can easily obtain a cross-lingual clustering. Figure 1 shows the complete algorithm. The intuition behind this method is that by creating the cross-lingual data in this way, we bias the clustering algorithm towards putting words that are translations of each other in the same cluster.

### 3.2 Treebank Lexicalization

We now describe how to introduce lexical representations  $\phi^{(l)}(x, y)$  to the model. Our approach is simple: we take the treebank data  $\mathcal{T}_1 \dots \mathcal{T}_m$  for the  $m$  source languages, together with the translation lexicons  $t(w, i, m + 1)$ . For any word  $w$  in the source treebank data, we can look up its translation  $t(w, i, m + 1)$  in the lexicon, and add this translated form to the underlying sentence. Features can now consider lexical identities derived in this way. In many cases the resulting translation will be the NULL word. However, we still make use of the representations  $\phi^{(p)}(x, y)$  and  $\phi^{(c)}(x, y)$  as previously defined; these representations are useful when the NULL translation is observed.

### 3.3 Integration with the Density-Driven Projection Method of Rasooli and Collins (2015)

In this section we describe a method for integrating our approach with the cross-lingual transfer method of Rasooli and Collins (2015), which makes use of density-driven projections.

In annotation projection methods (Hwa et al., 2005; McDonald et al., 2011), it is assumed that we have translation data  $\mathcal{B}_{i,j}$  for a source and target language, and that we have a dependency parser in the source language  $\mathcal{L}_i$ . The translation data consists of pairs  $(e, f)$  where  $e$  is a source language sentence, and  $f$  is a target language sentence. A method such as GIZA++ is used to derive an alignment between the words in  $e$  and  $f$ , for each sentence pair; the source language parser is used to

parse  $e$ . Each dependency in  $e$  is then potentially transferred through the alignments to create a dependency in the target sentence  $f$ . Once dependencies have been transferred in this way, a dependency parser can be trained on the dependencies in the target language.

The density-driven approach of Rasooli and Collins (2015) makes use of various definitions of “density” of the projected dependencies. For example,  $\mathcal{P}_{100}$  is the set of projected structures where the projected dependencies form a full projective parse tree for the sentence;  $\mathcal{P}_{80}$  is the set of projected structures where at least 80% of the words in the projected structure are a modifier in some dependency. An iterative training process is used, where the parsing algorithm is first trained on the set  $\mathcal{T}_{100}$  of complete structures, and where progressively less dense structures are introduced in learning.

We integrate our approach with the density-driven approach of Rasooli and Collins (2015) as follows: Consider the treebanks  $\mathcal{T}_1 \dots \mathcal{T}_m$  created using the lexicalization method of §3.2. We add all trees in these treebanks to the set  $\mathcal{P}_{100}$  of full trees used to initialize the method of Rasooli and Collins (2015). In addition we make use of the representations  $\phi^{(p)}$ ,  $\phi^{(c)}$  and  $\phi^{(l)}$ , throughout the learning process.

## 4 Experiments

This section first describes the experimental settings, then reports results.

### 4.1 Data and Tools

**Data** In a first set of experiments, we consider the 7 European languages studied in previous work (Ma and Xia, 2014; Zhang and Barzilay, 2015; Guo et al., 2016; Ammar et al., 2016; Lacroix et al., 2016). More specifically, we use the 7 European languages in the Google universal treebank (v.2; standard data) (McDonald et al., 2013). As in previous work, gold part-of-speech tags are used for evaluation. We use the concatenation of the treebank training sentences, Wikipedia data and the Bible monolingual sentences as our monolingual raw text. Table 3 shows statistics for the monolingual data. We use the Bible data from Christodouloupoulos and Steedman (2014), which includes data for 100 languages, as the source of translations. We also con-

duct experiments with the Europarl data (both with the original size and a subset of it with the same size as the Bible text) to study the effects of translation data size and domain shift. The statistics for translation data is shown in Table 4.

In a second set of experiments, we run experiments on 38 datasets (26 languages) in the more recent Universal Dependencies v1.3 corpus (Nivre et al., 2016).<sup>4</sup>

Lang.	en	de	es	fr	it	pt	sv
#Sen.	31.8	20.0	13.6	13.6	10.1	6.1	3.9
#Token	750.5	408.2	402.3	372.1	311.1	169.3	60.6
#Type	3.8	6.1	2.7	2.4	2.1	1.6	1.3

Table 3: Sizes of the monolingual datasets for each of our languages. All numbers are in millions.

**Brown Clustering Algorithm** We use the off-the-shelf Brown clustering tool<sup>5</sup> (Liang, 2005) to train monolingual Brown clusters with 500 clusters. The monolingual Brown clusters are used as features over lexicalized values created in  $\phi^{(l)}$ , and in self-training experiments. We train our cross-lingual clustering with the off-the-shelf-tool<sup>6</sup> from Stratos et al. (2015). We set the window size to 2 with cluster size of 500.<sup>7</sup>

**Parsing Model** We use the k-beam arc-eager dependency parser of Rasooli and Tetreault (2015), which is similar to the model of Zhang and Nivre (2011). We modify the parser such that it can use both monolingual and cross-lingual word cluster features. The parser is trained using the the maximum violation update strategy (Huang et al., 2012).

<sup>4</sup>We excluded languages that are not completely present in the Bible corpus of Christodouloupoulos and Steedman (2014) (Ancient Greek, Basque, Catalan, Galician, Gothic, Irish, Kazakh, Latvian, Old Church Slavonic, and Tamil). We also excluded Arabic, Hebrew, Japanese and Chinese, as these languages have tokenization and/or morphological complexity that goes beyond the scope of this paper. Future work should consider these languages.

<sup>5</sup><https://github.com/percyliang/brown-cluster>

<sup>6</sup><https://github.com/karlstratos/singular>

<sup>7</sup>Usually the original Brown clusters are better features for parsing but their training procedure does not scale on big datasets. Therefore we use the more efficient algorithm from Stratos et al. (2015) on the larger cross-lingual datasets to obtain word clusters.

Data	Lang.	en	de	es	fr	it	pt	sv
Bible	tokens	1.5M	665K	657K	732K	613K	670K	696K
	types	16K	20K	27K	22K	29K	29K	23K
EU-S	tokens	718K	686K	753K	799K	717K	739K	645K
	types	22K	41K	31K	27K	30K	32K	39K
Europarl	tokens	56M	50M	57M	62M	55M	56M	46M
	types	133K	400K	195K	153K	188K	200K	366K

Table 4: Statistics for the Bible, sampled Europarl (EU-S) and Europarl datasets. Each individual Bible text file from Christodouloupoulos and Steedman (2014) consists of 24720 sentences, except for English datasets, where two translations into English are available, giving double the amount of data. Each text file from the sampled Europarl datasets consists of 25K sentences and Europarl has approximately 2 million sentences per language pair.

We use three epochs of training for all experiments.

**Word alignment** We use the intersected alignments from Giza++ (Och and Ney, 2000) on translation data. We exclude sentences in translation data with more than 100 words.

## 4.2 Results on the Google Treebank

Table 5 shows the dependency parsing accuracy for the baseline delexicalized approach, and for models which add 1) cross-lingual clusters (§3.1); 2) lexical features (§3.2); 3) integration with the density-driven method of Rasooli and Collins (2015). It can be seen that each of these three steps gives significant improvements in performance. The final LAS/UAS of 73.90/80.28% is several percentage points higher than the baseline accuracy of 63.91/72.94%.

We also run experiments with a subset of the Europarl data to see the effect of domain shift. As shown in Table 5, we observe that there is a slight drop in accuracy when we use the Bible data. The main exception is the English language, where the number of translation sentences is twice more than the sampled Europarl sentences.

**Comparison to the Density-Driven Approach using Europarl Data** Table 6 shows accuracies for the density-driven approach of Rasooli and Collins

Lang.	Baseline		This paper using the Bible data					
			§3.1		§3.2		§3.3	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
en	58.20	65.50	65.04	72.35	66.33	74.00	<b>70.84</b>	<b>76.55</b>
de	49.72	59.15	51.59	59.71	54.90	62.65	<b>65.25</b>	<b>72.82</b>
es	68.29	77.18	73.12	79.63	76.60	81.87	<b>76.73</b>	<b>82.06</b>
fr	67.30	77.68	69.46	79.91	74.36	81.92	<b>75.78</b>	<b>82.22</b>
it	69.74	79.38	71.64	80.04	74.68	82.81	<b>76.15</b>	<b>83.30</b>
pt	71.48	77.47	76.90	81.54	81.02	84.44	<b>81.30</b>	<b>84.73</b>
sv	62.61	74.19	63.50	75.11	68.21	78.73	<b>71.24</b>	<b>80.30</b>
avg	63.91	72.94	67.32	75.47	70.87	78.06	<b>73.90</b>	<b>80.28</b>

Table 5: Performance of different models in this paper; first the baseline model, then models trained using the methods described in sections §3.1–3.3.

Lang.	Bible				Europarl-Sample				Europarl			
	Density		This Paper		Density		This Paper		Density		This Paper	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
en	59.13	66.36	70.84	76.55	64.28	72.81	70.16	76.18	68.44	76.27	71.07	77.49
de	60.21	69.53	65.25	72.82	61.58	72.03	64.90	73.01	73.01	79.68	75.57	82.13
es	70.29	76.85	76.73	82.06	72.04	78.34	76.03	81.55	74.61	80.86	76.64	82.56
fr	69.89	76.87	75.78	82.22	71.93	78.98	75.71	82.50	76.26	82.72	77.40	83.92
it	71.11	78.50	76.15	83.30	73.20	80.39	76.19	82.93	76.98	83.67	77.45	84.37
pt	72.07	76.41	81.30	84.73	75.32	79.69	81.61	84.83	77.30	82.07	82.13	85.61
sv	66.52	76.35	71.24	80.30	71.92	80.65	73.49	81.62	75.61	84.06	76.90	84.55
avg	67.03	75.75	73.90	80.28	70.04	77.56	74.01	80.37	74.60	81.33	76.74	82.95

Table 6: Results for our method using different sources of translation data. “Density” refers to the method of Rasooli and Collins (2015); “This paper” gives results using the methods described in sections 3.1–3.3 of this paper. The “Bible” experiments use the Bible data of Christodouloupoulos and Steedman (2014). The “Europarl” experiments use the Europarl data of Koehn (2005). The “Europarl-Sample” experiments use 25K randomly chosen sentences from Europarl; this gives a similar number of sentences to the Bible data.

Lang.	ZB15		GCY16		AMB16		RC15		This paper				Supervised	
									Bible		Europarl			
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS		
en	59.8	70.5	–	–	–	–	68.44	76.27	70.84	76.55	<b>71.07</b>	<b>77.49</b>	92.0	93.8
de	54.1	62.5	55.9	65.0	57.1	65.2	73.01	79.68	65.25	72.82	<b>75.57</b>	<b>82.13</b>	79.4	85.3
es	68.3	78.0	73.0	79.0	74.6	80.2	74.61	80.86	<b>76.73</b>	82.06	76.03	<b>82.56</b>	82.3	86.7
fr	68.8	78.9	71.0	77.6	73.9	80.6	76.26	82.72	75.78	82.22	<b>77.40</b>	<b>83.92</b>	81.7	86.3
it	69.4	79.3	71.2	78.4	72.5	80.7	76.98	83.67	76.15	83.30	<b>77.45</b>	<b>84.37</b>	86.1	88.8
pt	72.5	78.6	78.6	81.8	77.0	81.2	77.30	82.07	81.30	84.73	<b>82.13</b>	<b>85.61</b>	87.6	89.4
sv	62.5	75.0	69.5	78.2	68.1	79.0	75.61	84.06	71.24	80.30	<b>76.90</b>	<b>84.55</b>	84.1	88.1
avg <sub>en</sub>	65.9	75.4	69.3	76.3	70.5	77.8	75.63	82.18	74.41	80.91	<b>77.68</b>	<b>83.86</b>	83.5	87.4

Table 7: Comparison of our work using Bible and Europarl data, with previous work: ZB15 (Zhang and Barzilay, 2015), GCY16 (Guo et al., 2016), AMB 16 (Ammar et al., 2016), and RC15 (Rasooli and Collins, 2015). “Supervised” refers to the performance of the parser trained on fully gold standard data in a supervised fashion (i.e. the practical upper-bound of our model). “avg<sub>en</sub>” refers to the average accuracy for all datasets except English.

(2015), first using Europarl data<sup>8</sup> and second using the Bible data alone (with no cross-lingual clusters or lexicalization). The Bible data is considerably smaller than Europarl (around 100 times smaller), and it can be seen that results using the Bible are several percentage points lower than the results for Europarl (75.75% UAS vs. 81.33% UAS). Integrating cluster and lexicalized features described in the current paper with the density-driven approach closes much of this gap in performance (80.28% UAS). Thus we have demonstrated that we can get close to the performance of the Europarl-based models using only the Bible as a source of translation data. Using our approach on the full Europarl data gives an average UAS of 82.95%, an improvement from the 81.33% UAS of Rasooli and Collins (2015).

Table 6 also shows results when we use a random subset of the Europarl data, where the number of sentences (25,000) is chosen to give a very similar size to the Bible dataset. It can be seen that accuracies using the Bible data vs. Europarl-Sample are very similar (80.28% vs. 80.37% UAS), suggesting that the size of the translation corpus is much more important than the genre.

**Comparison to Other Previous Work** Table 7 compares the accuracy of our method to the following related work: 1) Zhang and Barzilay (2015), who describe a method that learns cross-lingual embeddings and bilingual dictionaries from Europarl data, and uses these features in a discriminative parsing model; 2) Guo et al. (2016), who describe a method that learns cross-lingual embeddings from Europarl data and uses a shift-reduce neural parser with these representations; 3) Ammar et al. (2016), who use the same embeddings as Guo et al. (2016), within an LSTM-based parser; and 4) Rasooli and Collins (2015) who use the density-driven approach on the Europarl data. Our method gives significant improvements over the first three models, in spite of using the Bible translation data rather than Europarl. When using the Europarl data, our method improves the state-of-the-art model of Rasooli and Collins (2015). It is worth noting that our model significantly outperform the UAS reported in other

<sup>8</sup>Rasooli and Collins (2015) do not report results on English. We use the same setting as in their paper to obtain the English results.

Lang.	RC15		This Paper (§3.3)			
			Bible		Europarl	
	LAS	UAS	LAS	UAS	LAS	UAS
en	66.23	74.44	67.76	74.44	<b>68.00</b>	<b>75.15</b>
de	71.57	78.77	61.86	70.33	<b>73.58</b>	<b>80.80</b>
es	72.33	79.17	73.78	79.86	<b>74.17</b>	<b>80.74</b>
fr	73.51	80.77	72.62	79.92	<b>75.00</b>	<b>82.32</b>
it	74.91	82.03	74.01	81.68	<b>75.34</b>	<b>82.64</b>
pt	75.4	80.67	79.22	83.31	<b>80.45</b>	<b>84.41</b>
sv	73.42	82.03	67.29	77.17	<b>73.72</b>	<b>82.25</b>
avg	72.48	79.70	70.93	78.10	<b>74.32</b>	<b>81.19</b>

Table 8: The final results based on automatic part of speech tags. RC15 refers to the best performing model of Rasooli and Collins (2015).

previous work (McDonald et al., 2011; Ma and Xia, 2014; Lacroix et al., 2016) but we do not put theirs because of space restrictions.

**Performance with Automatic POS Tags** For completeness, Table 8 gives results for our method with automatic part-of-speech tags. Although there is a drop in accuracy when using the automatic POS tags, our model has still a high accuracy even higher than those from the density driven approach of Rasooli and Collins (2015).

### 4.3 Results on the Universal Dependencies v1.3

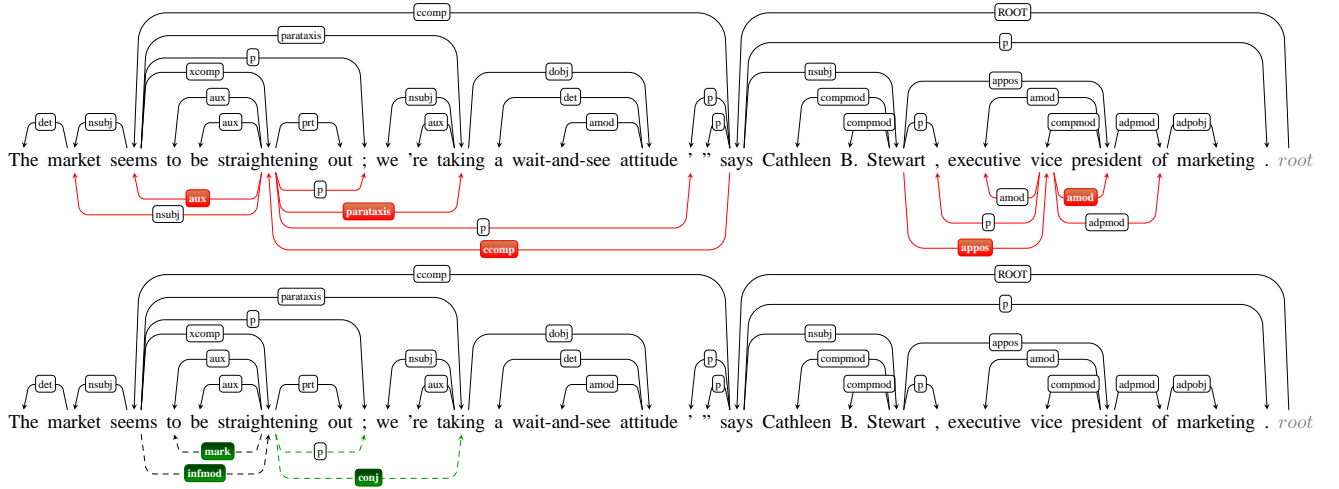
Table 9 gives results on 38 datasets (26 languages) from the newly released universal dependencies corpus (Nivre et al., 2016). Given the number of treebanks and to speed up training, we pick source languages that have at least 5 out of 6 common WALS properties with each target language. Our experiments are carried out using the Bible as our translation data. As shown in Table 9, our method consistently outperforms the density-driven method of Rasooli and Collins (2015) and for many languages the accuracy of our method gets close to the accuracy of the supervised parser. Accuracy on some languages (e.g., Persian (fa) and Turkish (tr)) is low, suggesting that future work should consider more powerful techniques for these languages.

## 5 Analysis

We conclude this paper with some examples from the English development data (Figures 3 and 4). We



(a)



(b)

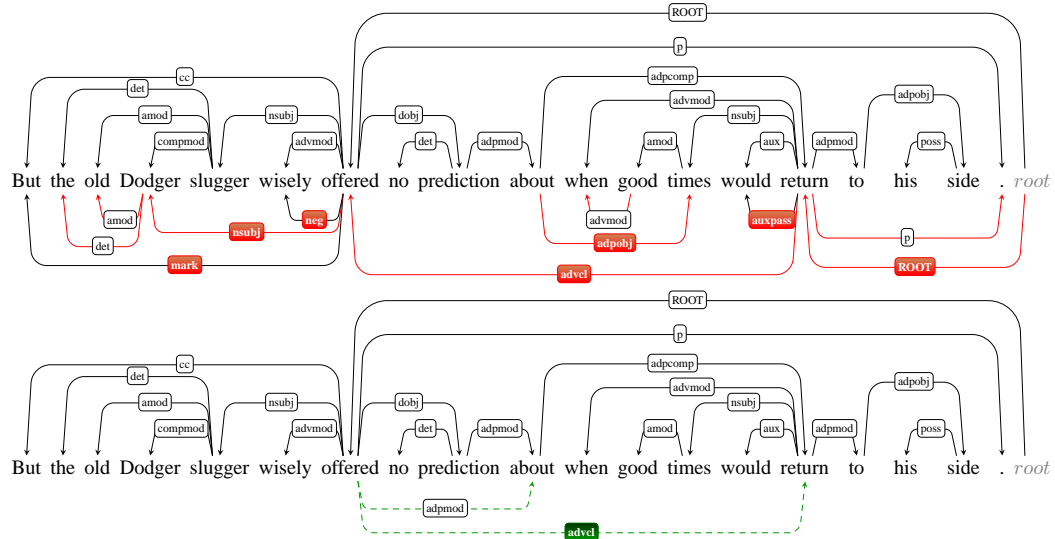


Figure 3: Two English sentences from the development data for which our method with the Europarl data correctly recovers the full tree, but the baseline model and the Bible-based model have some errors. The correct dependency parse is shown above each sentence. Incorrect dependencies from the baseline model are shown with solid red edges (top figure in each case) and incorrect dependencies from the Bible model are shown with dashed green edges (bottom figure in each case).

find interesting examples for which our method predicts the correct tree, including some long sentences (Figures 3a and 3b). Except for very short sentences with basic grammatical structures, the baseline parser completely fails to recover the correct structure. In Figure 2, two examples are depicted for which the baseline parser has a very low accuracy while our method is fully accurate. In the first

example, the baseline parser attaches the root node to a wrong verb (“plug” instead of “goes”) and subsequently attaches incorrect dependents. In the first example in Figure 2, the baseline parser cannot find the head of the noun phrase (“diversified Fidelity funds”). Because of the lack of lexical features, the baseline parser attaches the word “after” as a prepositional phrase instead of assigning it as a dependent



Dataset	Density		This paper		Supervised	
	LAS	UAS	LAS	UAS	LAS	UAS
it	74.3	81.3	79.8	86.1	88.4	90.7
sl	68.2	75.9	78.6	84.1	86.3	89.1
es	69.1	77.5	76.3	84.1	83.5	86.9
bg	66.2	79.5	72.0	83.6	85.5	90.5
pt	66.7	75.8	74.8	83.4	83.0	86.7
es-ancora	68.9	77.5	74.6	83.1	86.5	89.4
fr	72.0	77.9	76.6	82.6	84.5	87.1
sv-lines	67.5	76.7	73.3	82.4	81.0	85.4
pt-br	68.3	75.2	76.2	82.0	87.8	89.7
sv	65.9	75.7	71.7	81.3	83.6	87.7
no	71.7	78.8	74.3	81.2	88.0	90.5
pl	65.4	77.6	70.1	81.0	85.1	90.3
hr	55.8	70.2	65.9	80.9	76.2	85.1
cs-cac	61.1	70.3	69.0	78.5	82.4	87.6
da	63.1	72.8	68.3	77.8	80.8	84.3
en-lines	67.0	75.9	68.6	77.3	80.7	84.6
cs	59.0	68.1	67.2	76.4	84.5	88.7
id	38.0	55.7	57.8	76.0	79.8	85.1
de	61.3	72.8	64.9	75.7	80.2	85.8
ru-syntagrus	56.0	70.7	61.6	75.3	82.0	87.8
ru	56.7	64.8	65.4	74.8	71.9	77.7
cs-cltt	57.5	65.4	65.6	74.7	77.1	81.4
ro	54.6	67.4	60.7	74.6	78.2	85.3
la	54.5	71.6	55.7	72.8	43.1	52.5
nl-lassysmall	51.5	62.6	61.9	71.7	76.5	80.6
el	53.7	66.7	59.6	71.0	79.1	83.1
et	48.9	65.6	56.9	70.9	75.9	82.9
hi	34.4	50.6	49.9	69.9	89.4	92.9
hu	26.1	48.9	55.0	69.9	69.5	79.4
en	59.7	68.1	61.8	69.0	85.3	88.1
fi-ftb	50.3	63.2	56.5	67.5	73.3	79.7
fi	49.8	60.8	57.3	66.4	73.4	78.2
la-ittb	44.1	55.4	51.8	62.8	76.2	80.9
nl	40.6	49.4	50.1	62.0	70.1	75.0
la-proiel	43.6	60.3	45.0	61.3	64.9	72.9
sl-sst	42.4	59.2	47.6	60.6	63.4	70.4
fa	44.4	53.2	46.5	56.0	84.1	87.5
tr	05.3	18.5	32.7	51.9	65.6	78.8
Average	56.7	68.1	64.0	74.8	78.9	83.8

Table 9: Results for the density driven method (Rasooli and Collins, 2015) and ours using the Bible data on the universal dependencies v1.3 (Nivre et al., 2016). The table is sorted by the performance of our method. The last major columns shows the performance of the supervised parser. The abbreviations are as follows: bg (Bulgarian), cs (Czech), da (Danish), de (German), el (Greek), en (English), es (Spanish), et (Estonian), fa (Persian (Farsi)), fi (Finnish), fr (French), hi (Hindi), hr (Croatian), hu (Hungarian), id (Indonesian), it (Italian), la (Latin), nl (Dutch), no (Norwegian), pl (Polish), pt (Portuguese), ro (Romanian), ru (Russian), sl (Slovenian), sv (Swedish), and tr (Turkish).

has introduced cross-lingual representations –for example cross-lingual word-embeddings– that can be used to improve performance (Zhang and Barzilay, 2015; Guo et al., 2015; Duong et al., 2015; Guo et al., 2016; Ammar et al., 2016). These cross-lingual representations are usually learned from parallel translation data. We show results for the methods of (Zhang and Barzilay, 2015; Guo et al., 2016; Ammar et al., 2016) in Table 7 of this paper.

The annotation projection approach, where dependencies from one language are transferred through translation alignments to another language, has been considered by several authors (Hwa et al., 2005; Ganchev et al., 2009; McDonald et al., 2011; Ma and Xia, 2014; Rasooli and Collins, 2015; Lacroix et al., 2016; Agić et al., 2016). The work of Rasooli and Collins (2015) gives the best results of these methods on the Europarl data, as shown in Table 7.

Other recent work (Tiedemann et al., 2014; Tiedemann, 2015; Tiedemann and Agić, 2016) has considered treebank translation, where a statistical machine translation system (e.g., MOSES (Koehn et al., 2007)) is used to translate a source language treebank into the target language, complete with re-ordering of the input sentence. The lexicalization approach described in this paper is a simple form of treebank translation, where we use a word-to-word translation model. In spite of its simplicity, it is a very effective approach.

A number of authors have considered incorporating universal syntactic properties such as dependency order by selectively learning syntactic attributes from similar source languages (Naseem et al., 2012; Täckström et al., 2013; Zhang and Barzilay, 2015; Ammar et al., 2016). Selective sharing of syntactic properties is complementary to our work; we used a very limited form of selective sharing, through the WALS properties, in our baseline approach.

A number of authors (Täckström et al., 2012; Guo et al., 2015; Guo et al., 2016) have introduced methods that learn cross-lingual representations that are then used in syntactic transfer. Most of these approaches introduce constraints to a clustering or embedding algorithm that encourage words that are translations of each other to have similar representations. Our method of deriving a cross-lingual corpus

(see Figure 1) which is then used as input to a clustering algorithm is closely related to (Duong et al., 2015; Gouws and Søgaard, 2015; Wick et al., 2015).

## 7 Conclusions

We have described a method for cross-lingual syntactic transfer that is effective in the scenario where a large amount of translation data is not available. We have introduced a simple, direct method for deriving cross-lingual clusters, and for transferring lexical information across treebanks for different languages. Experiments with the method show that the method gives improved performance over previous work that makes use of Europarl, a much larger translation corpus.

## Appendix A Parsing Features

We used all features in (Zhang and Nivre, 2011, Table 1 and 2), which describes features based on the word and part-of-speech at various positions on the stack and buffer of the transition system. In addition, we expand the Zhang and Nivre (2011, Table 1) features to include clusters, as follows: whenever a feature tests the part-of-speech for a word in position 0 of the stack or buffer, we introduce features that replace the part-of-speech with the Brown clustering bit-string of length 4 and 6. Whenever a feature tests for the word identity at position 0 of the stack or buffer, we introduce a cluster feature that replaces the word with the full cluster feature. We take the cross product of all features corresponding to the choice of 4 or 6 length bit string for part-of-speech features.

## References

- Željko Agić, Anders Johannsen, Barbara Plank, Héctor Alonso Martínez, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.
- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. One parser, many languages. *arXiv preprint arXiv:1602.01595*.
- Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jennifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.
- Christos Christodouloupoulos and Mark Steedman. 2014. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, pages 1–21.
- Shay B. Cohen, Dipanjan Das, and Noah A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 50–61, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics, July.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Cross-lingual transfer for unsupervised dependency parsing without parallel data. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 113–122, Beijing, China, July. Association for Computational Linguistics.
- Greg Durrett, Adam Pauls, and Dan Klein. 2012. Syntactic transfer using a bilingual lexicon. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–11, Jeju Island, Korea, July. Association for Computational Linguistics.
- Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 369–377, Suntec, Singapore, August. Association for Computational Linguistics.
- Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado, May–June. Association for Computational Linguistics.
- Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244,

- Beijing, China, July. Association for Computational Linguistics.
- Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *The Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, Phoenix, Arizona, USA.
- Liang Huang, Suphan Fayong, and Yang Guo. 2012. Structured perceptron with inexact search. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–151, Montréal, Canada, June. Association for Computational Linguistics.
- Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering*, 11(03):311–325.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.
- Ophélie Lacroix, Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1058–1063, San Diego, California, June. Association for Computational Linguistics.
- Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.
- Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1337–1348, Baltimore, Maryland, June. Association for Computational Linguistics.
- David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL '06*, pages 152–159, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 629–637. Association for Computational Linguistics.
- Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Çar Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Drogonova, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gokirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Simon Krek, Veronika Laippala, Lucia Lam, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Măranduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Loganathan Ramasamy, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji,

- Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uriu, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jing Xian Wang, Jonathan North Washington, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2016. Universal dependencies 1.3. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.
- Franz Josef Och and Hermann Ney. 2000. Giza++: Training of statistical translation models.
- Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338, Lisbon, Portugal, September. Association for Computational Linguistics.
- Mohammad Sadegh Rasooli and Joel Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *arXiv preprint arXiv:1503.06733*.
- Rudolf Rosa and Zdenek Zabokrtsky. 2015. Klcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 243–249, Beijing, China, July. Association for Computational Linguistics.
- Karl Stratos, Michael Collins, and Daniel Hsu. 2015. Model-based word embeddings from decompositions of count matrices. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1282–1291, Beijing, China, July. Association for Computational Linguistics.
- Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 477–487. Association for Computational Linguistics.
- Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. *Transactions for ACL*.
- Jörg Tiedemann and Željko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research*, 55:209–248.
- Jörg Tiedemann, Željko Agić, and Joakim Nivre. 2014. Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Jörg Tiedemann. 2015. Improving the cross-lingual projection of syntactic dependencies. In *Nordic Conference of Computational Linguistics NODALIDA 2015*, pages 191–199.
- Michael Wick, Pallika Kanani, and Adam Pocock. 2015. Minimally-constrained multilingual embeddings via artificial code-switching. In *Workshop on Transfer and Multi-Task Learning: Trends and New Perspectives*, Montreal, Canada, December.
- Min Xiao and Yuhong Guo. 2015. Annotation projection-based representation learning for cross-lingual dependency parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 73–82, Beijing, China, July. Association for Computational Linguistics.
- Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.
- Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, September.
- Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA, June. Association for Computational Linguistics.